Amazon Web Services
Data Engineering Immersion Day

Lab 1. Hydrating the Data Lake with DMS

Table of Contents

Introduction

This lab will give you an understanding of the AWS Database Migration Service (AWS DMS). You will migrate data from an existing Amazon Relational Database Service (Amazon RDS) Postgres database to an Amazon Simple Storage Service (Amazon S3) bucket that you create.

In this lab you will complete the following tasks:

1. Create a subnet group within the DMS Lab VPC
2. Create a DMS replication instance
3. Create a source endpoint
4. Create a target endpoint
5. Create a task to perform the initial migration of the data.

Optionally, you can add ongoing replication of data changes on the source

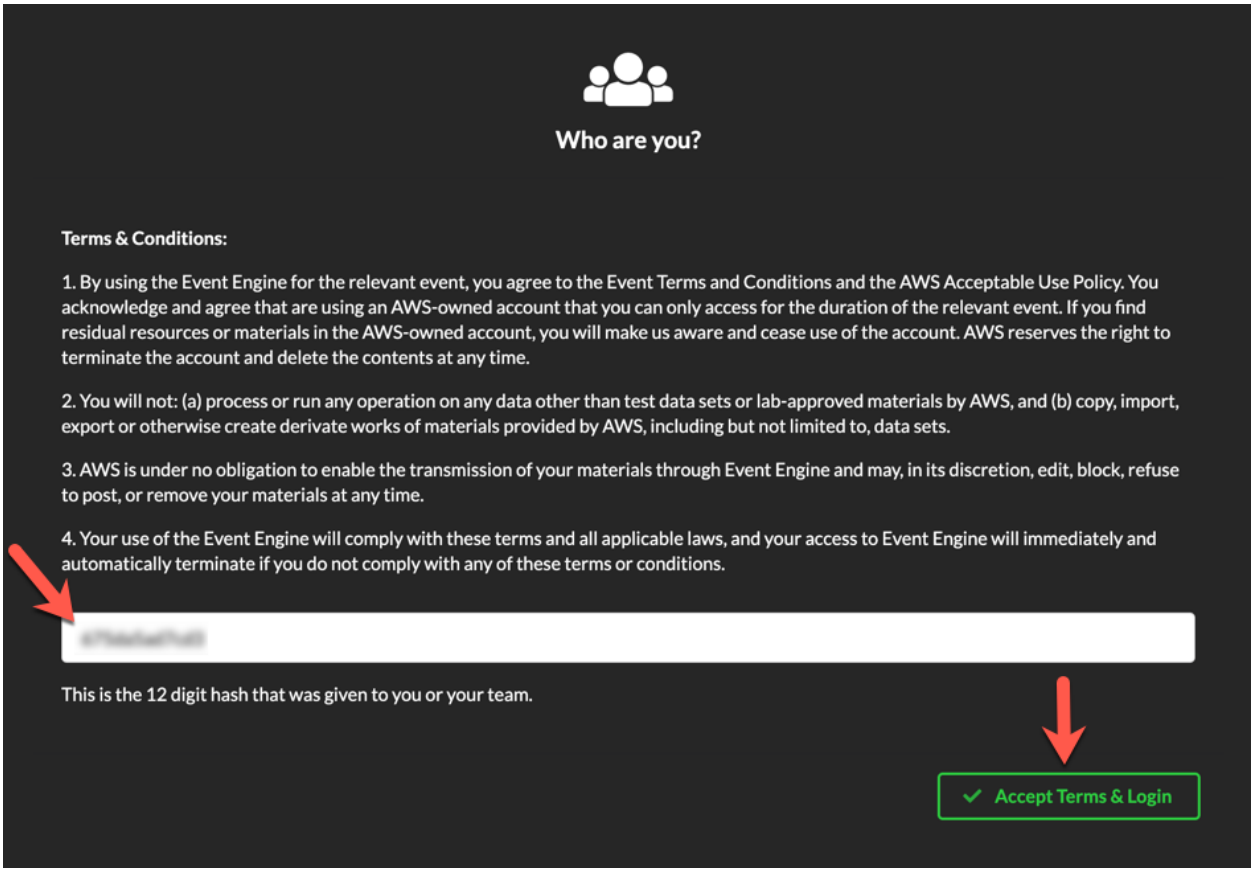
6. Create target endpoint for CDC files to place these files in a separate location than the initial load files
7. Create a task to perform the ongoing replication of data changes

Get Started Using the Lab Environment

Today, you are attending a formal event and you will have been sent your access details beforehand. If in the future you might want to perform these labs in your own AWS environment by yourself, you can follow instructions on GitHub - https://github.com/aws-samples/data-engineering-for-aws-immersion-day.

A 12-character access code (or 'hash') is the access code that grants you permission to use a dedicated AWS account for the purposes of this workshop.

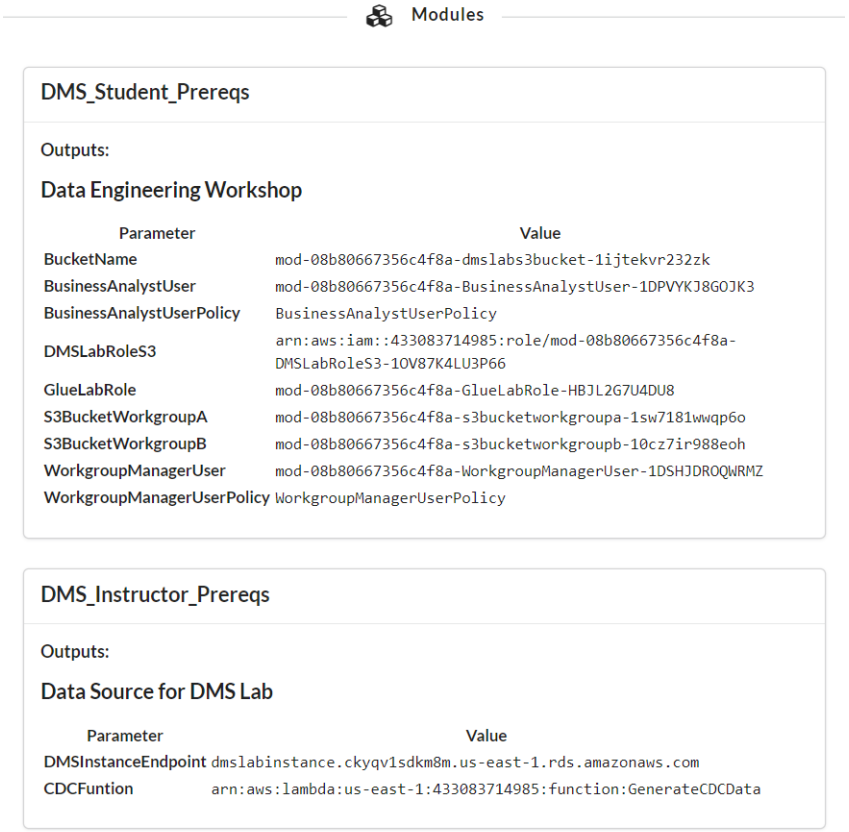
1. Go to https://dashboard.eventengine.run/, enter the access code and click Proceed:



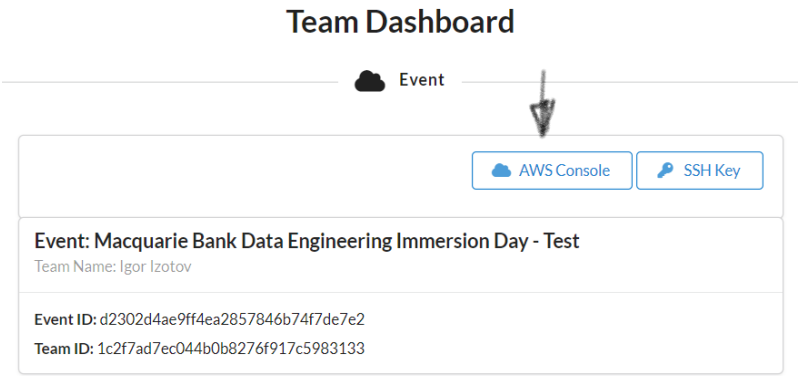
2. On the Team Dashboard web page you will see a set of connection strings and parameters that you will need during the labs. Best to save them to a text file locally, alternatively you can always go to this page to review them. Replace the parameters with the corresponding values from here where indicated in subsequent labs:

   Because you're at a formal event, some AWS resources have been pre-deployed for your convenience, for example

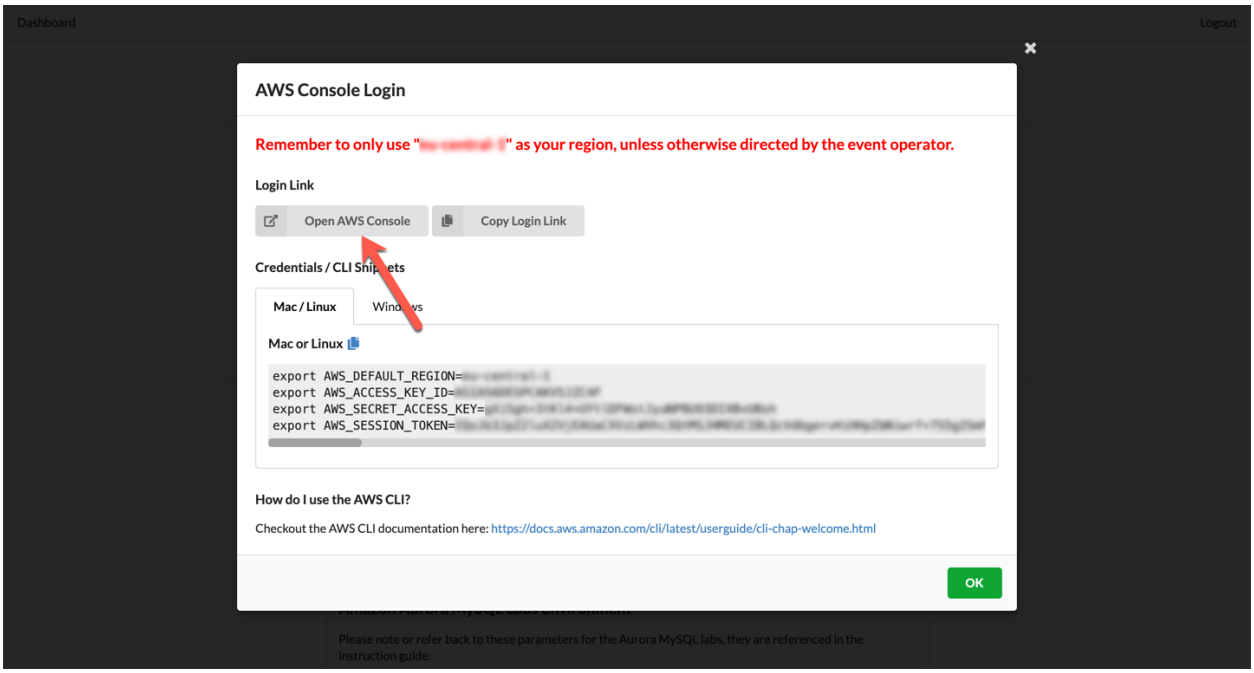
- RDS Postgres database that you will use as your source endpoint (parameter **`DMSInstanceEndpoint`**)
- VPC that you will deploy DMS to, etc



3. On the Team Dashboard, please click AWS Console to log into the AWS Management Console:



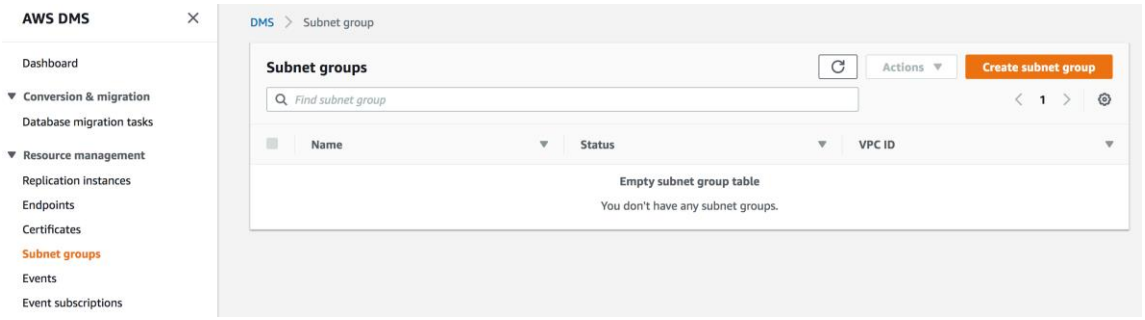
4. Click Open Console. For the purposes of this workshop, you will not need to use command line and API access credentials:



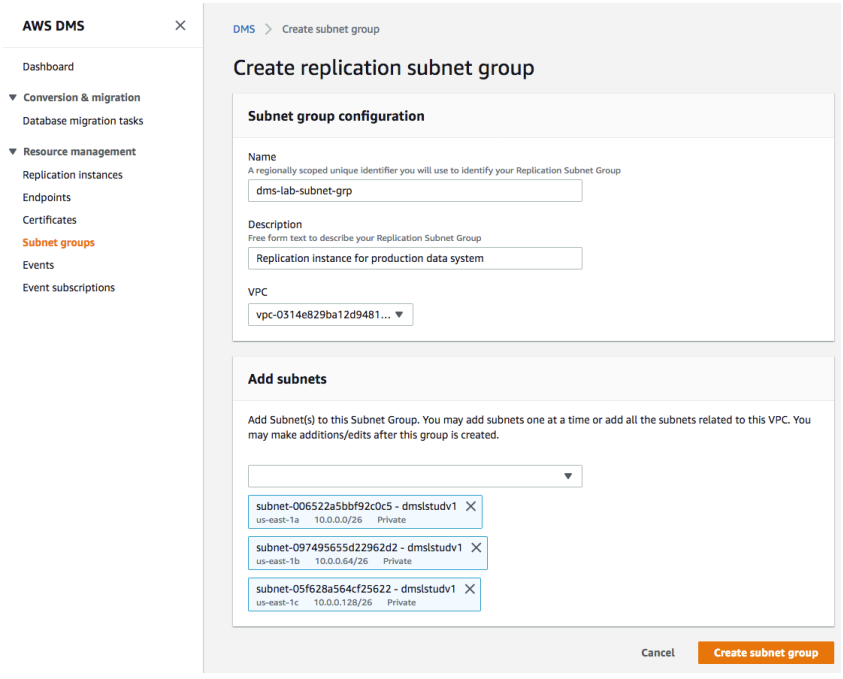
Once you have completed these steps, you can continue with the rest of this lab

Create the Subnet Group

1. Navigate to the DMS Console: https://console.aws.amazon.com/dms/v2/home?region=us-east-1#firstRun
2. On the DMS console, select **Subnet Groups**.

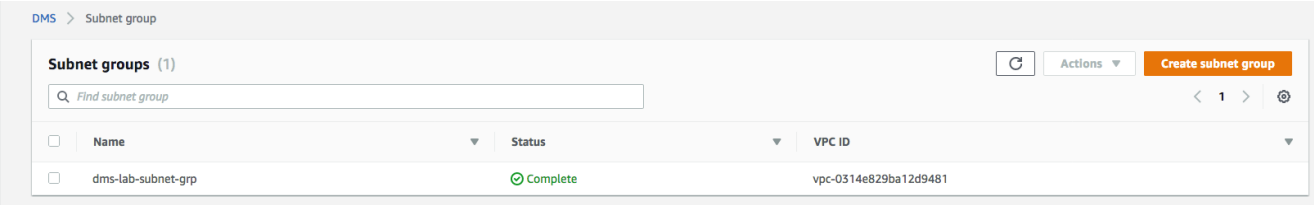


3. Click **Create subnet group**.
    a. In the Identifier box, type a descriptive name that you will easily recognize (e.g., “dms-lab-subnet-grp”).
    b. In the Description box, type an easily recognizable description (e.g., “**Replication instance for production data system**”).
    c. For VPC, select the pre-created VPC ending with **dmslstudv1**.
    The subnet list populates in the Available Subnets pane.
    d. Select as many subnets as you want and click Add. The selected subnets move to the Subnet Group pane. Note: DMS requires at least two separate availability zones to be selected.



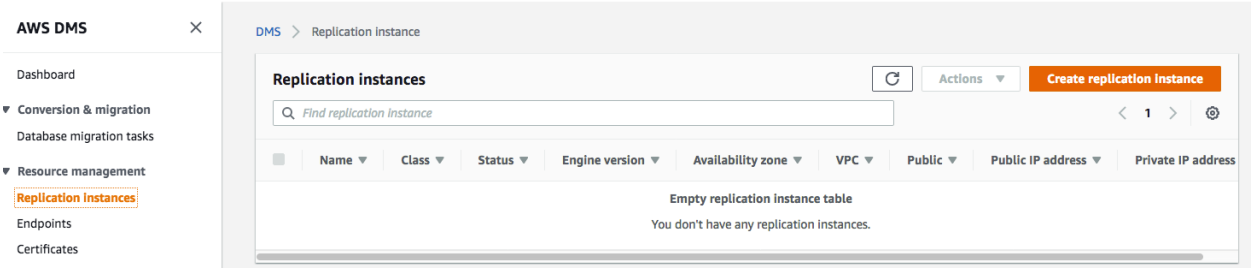
4. Click Create subnet group

5. On the DMS console, the subnet group status displays Complete.



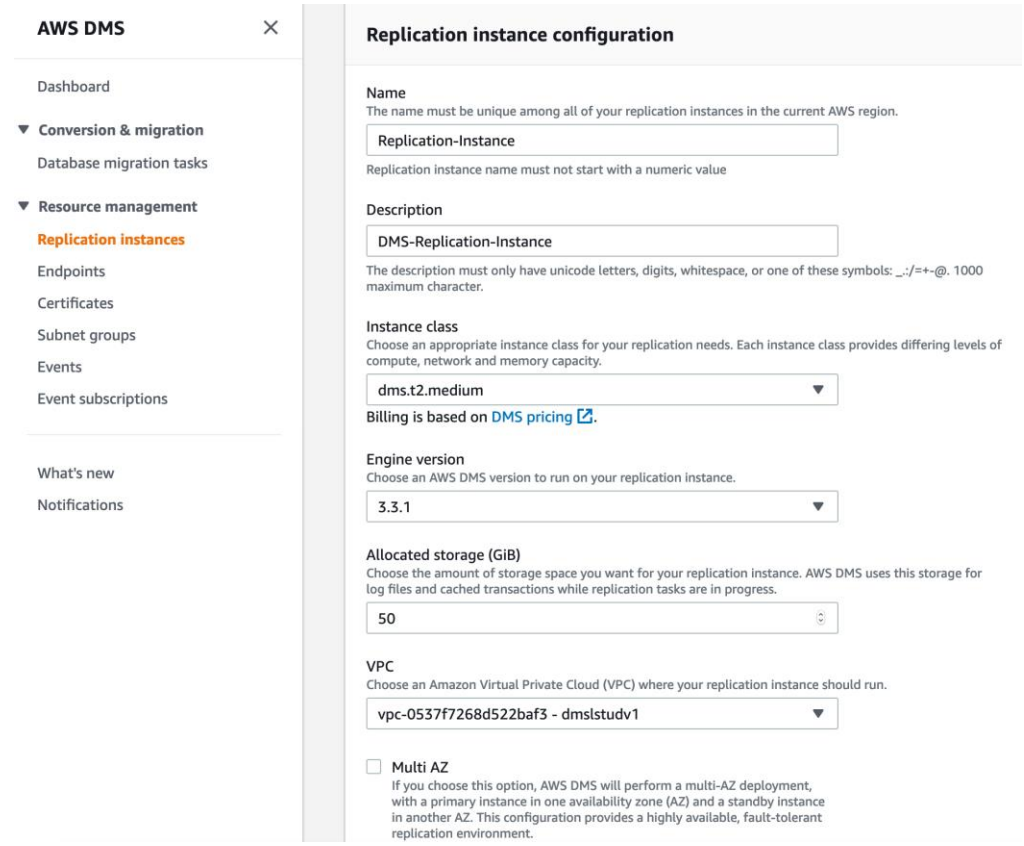
Create the Replication Instance

1. On the DMS console, select **Replication instances**.



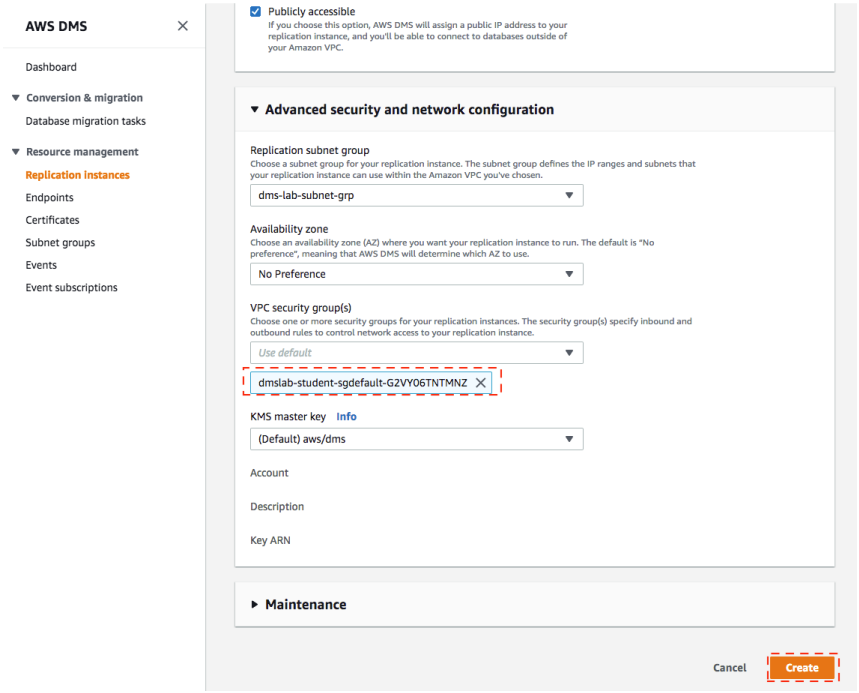
2. Click **Create replication instance**.

    a. For **Name**, type a name for the replication instance that you will easily recognize. (e.g., “DMS-Replication-Instance”).
    b. For **Description**, type a description you will easily recognize. (e.g., “DMS Replication Instance”).
    c. For **Instance class**, choose **dms.t2.medium**
    d. Select **Engine version** as **3.3.1**
    e. For VPC, select the name of the VPC that you created earlier with AWS CloudFormation template. VPC name ending with **dmslstudv1**



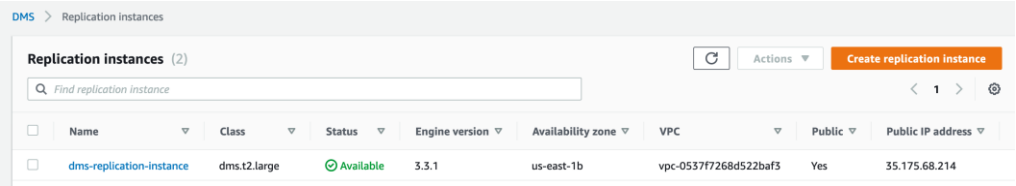
f. Click **Advanced** to expand the section.

g. Select the security group with **sgdefault** in the name.



3. Click **Create**.

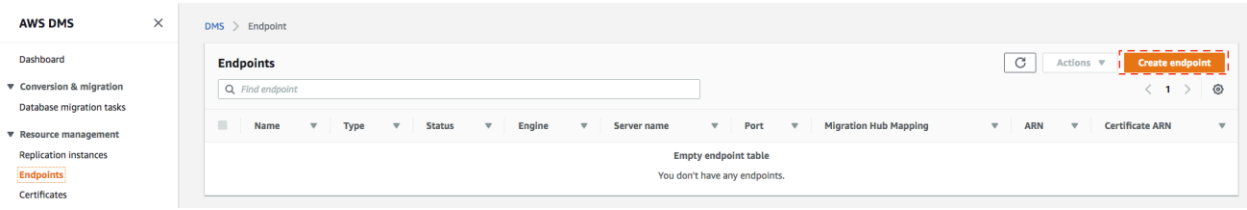
4. The DMS console displays **creating** for the instance status. When the replication instance is ready, the status changes to **available**. While replication instance is spinning up, you can proceed to next step for DMS endpoint creation.



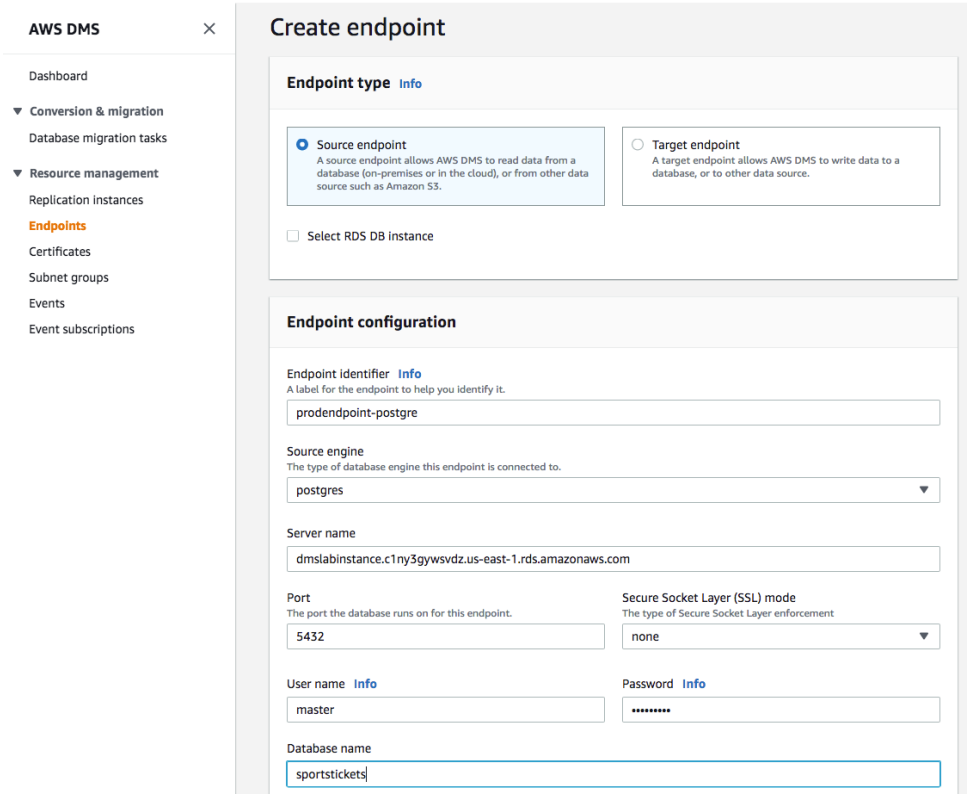
Create the DMS Source Endpoint

Please proceed to create your endpoints, without waiting for the step above.

1. On the DMS console, select **Endpoints**

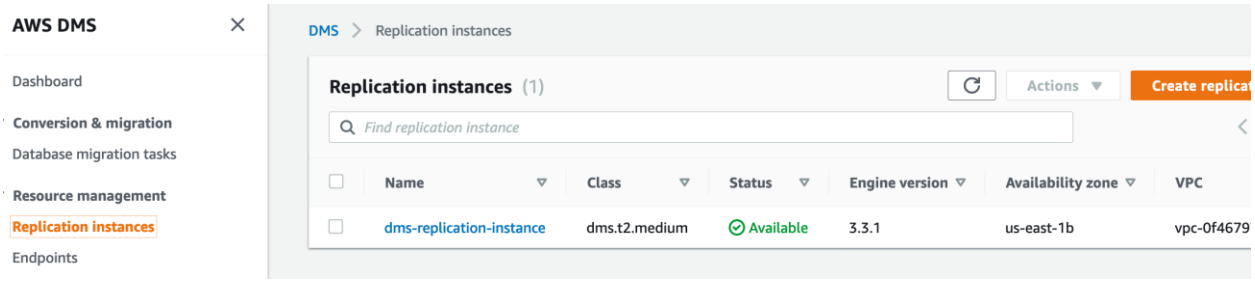


2. Click **Create endpoint**.
    a. select **Source endpoint** type.
    b. For **Endpoint identifier**, select an easily recognizable name (e.g. **rds-source-endpoint**)
    c. For Source engine, select **postgres**
    d. For Server, enter the value of **DMSInstanceEndpoint** parameter value from your Team Dashboard
    e. For Port, enter **5432**
    f. For SSL mode, choose **none**
    g. For User name, type **master**
    h. For Password, type **master123**
    i. For Database name, type **sportstickets**

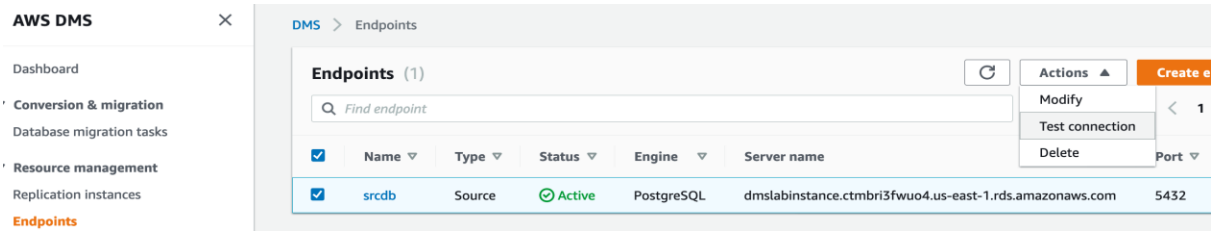


3. Click **Create endpoint** to create the endpoint. When available, the endpoint status changes to **active**.

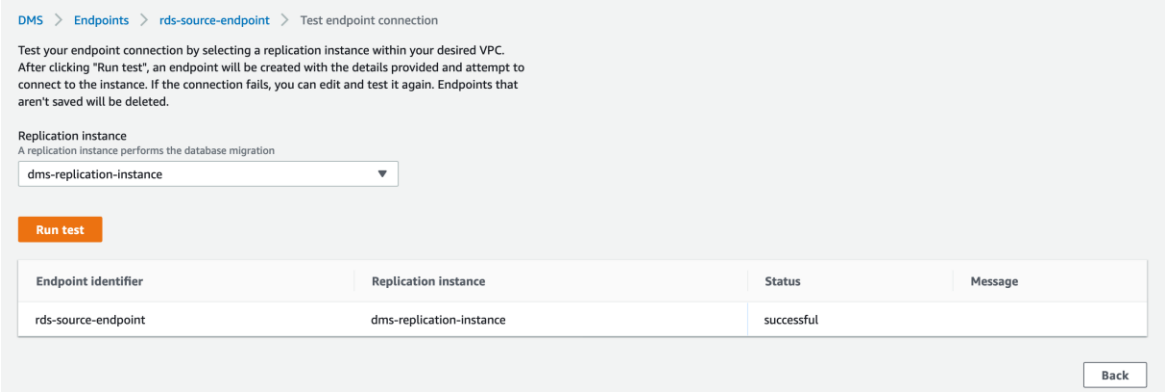
4. Check the **replication instance** created previously. Make sure the status is **available**.



5. Select your newly created source **endpoint**, and choose **Test connection** on the **Actions** drop-down list.



6. Click **Run test**. This step tests connectivity to the source database system. If successful, the message “Connection tested successfully” appears. **You may need to wait for the DMS replication instance to become available first.**

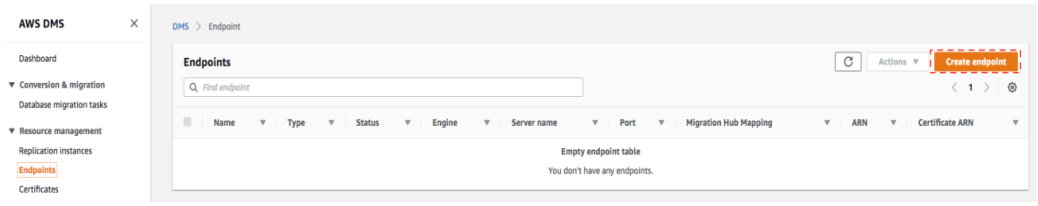


Create the Target Endpoint

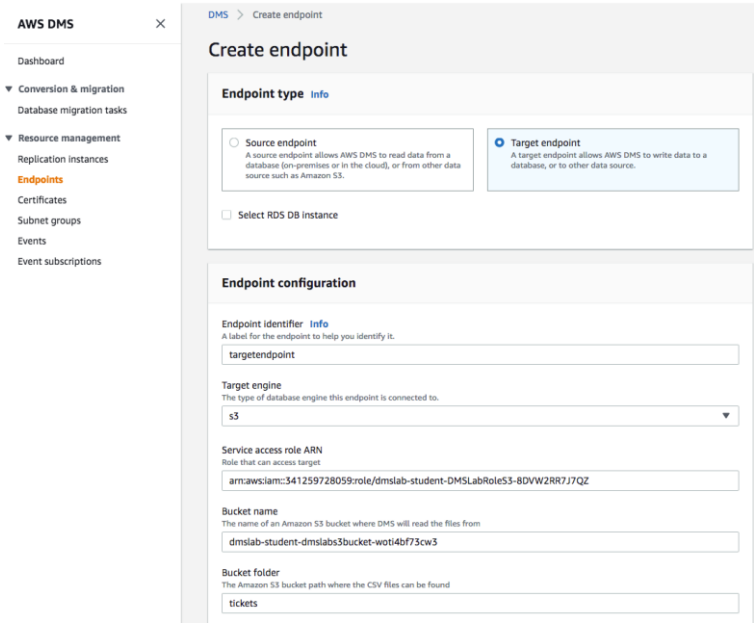
Before start, make sure you have the following values handy (can be found on your team dashboard):

- **DMSLabRoleS3** ARN – It looks like “arn:aws:iam::<Account number>:role/xxx-DMSLabRoleS3-xxxx“
- **BucketName** - It looks like “xxx-dmslabs3bucket-xxxx“

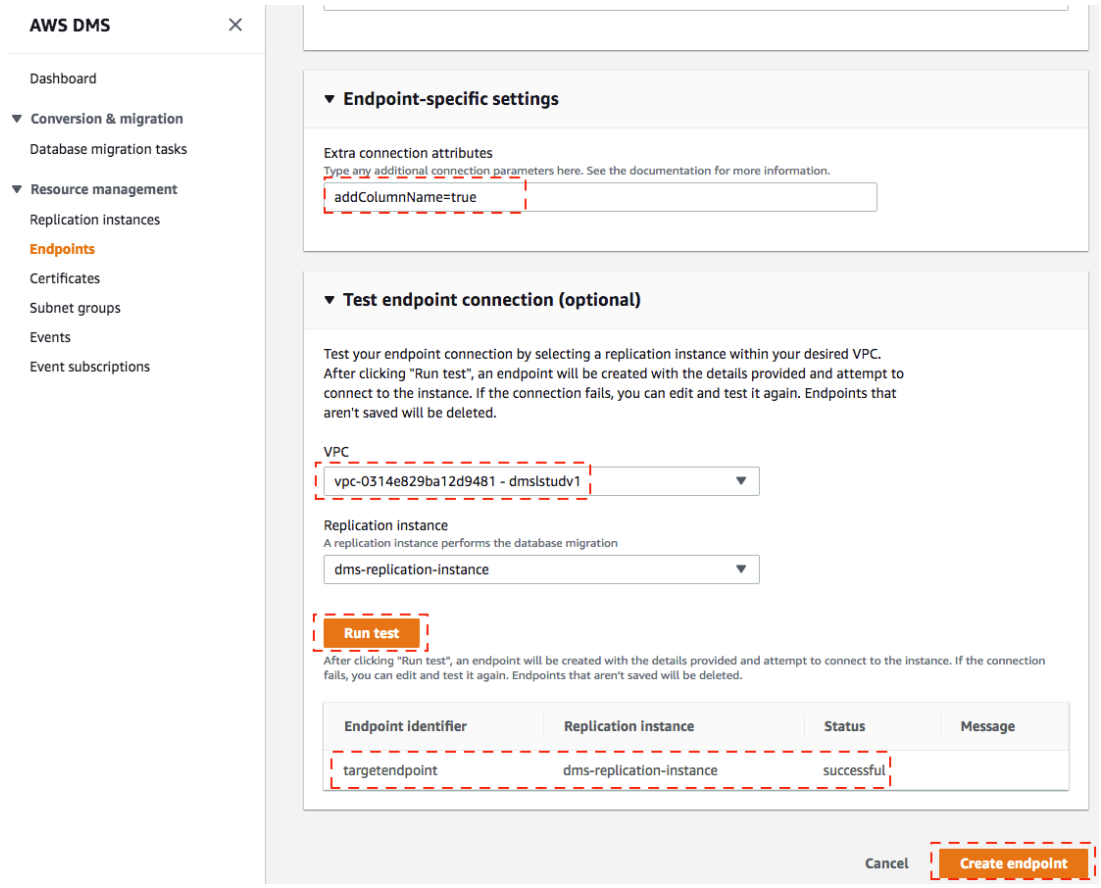
1. On the DMS console, select **Endpoints**.



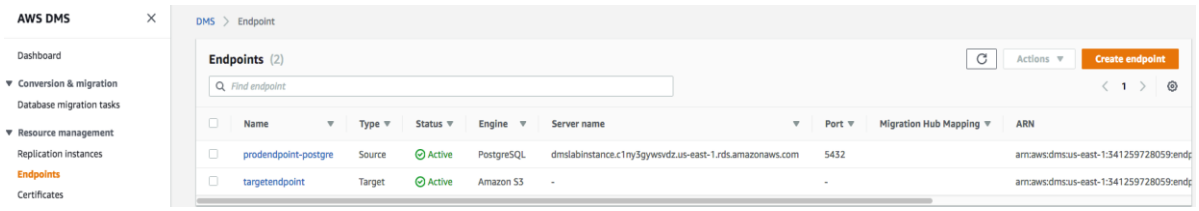
2. Click **Create endpoint**.
    a. For Endpoint type, select **Target endpoint**.
    b. For Endpoint identifier, type an easily recognized name such as **s3-target-endpoint**.
    c. For Target engine, choose **s3**.
    d. For Service access role ARN, paste the **DMSLabRoleS3** value noted earlier
    e. For Bucket name, paste the value of **BucketName** noted earlier
    f. For Bucket folder, type **tickets**



g. Click **Endpoint-specific settings** to expand the section.
h. In the **Extra connection attributes** box, type **addColumnName=true**. This attribute includes the column names in the files in the S3 bucket.
i. Expand the **Test endpoint connection (optional) section**, and choose your VPC name with **dmslstudv1** on the VPC drop-down list.
j. Click **Run test**. This step tests connectivity to the source database system. If successful, the message “Connection tested successfully” appears.

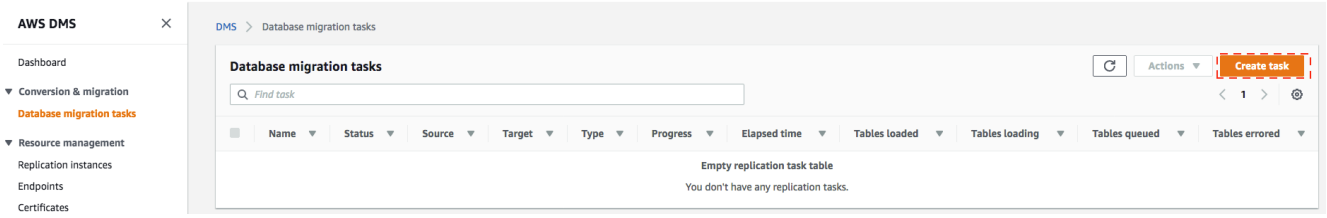


3. Click **Create Endpoint**. When available, the endpoint status changes to **active**.



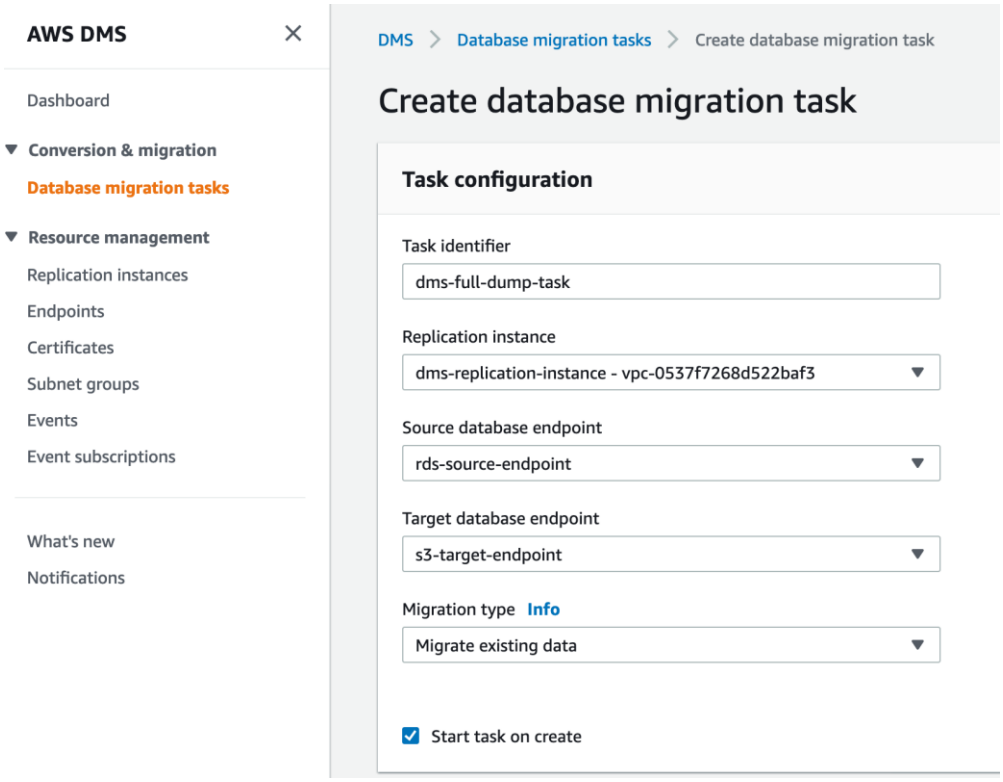
Create a task to perform the initial full copy

1. On the DMS console, select **Database Migration Tasks**.

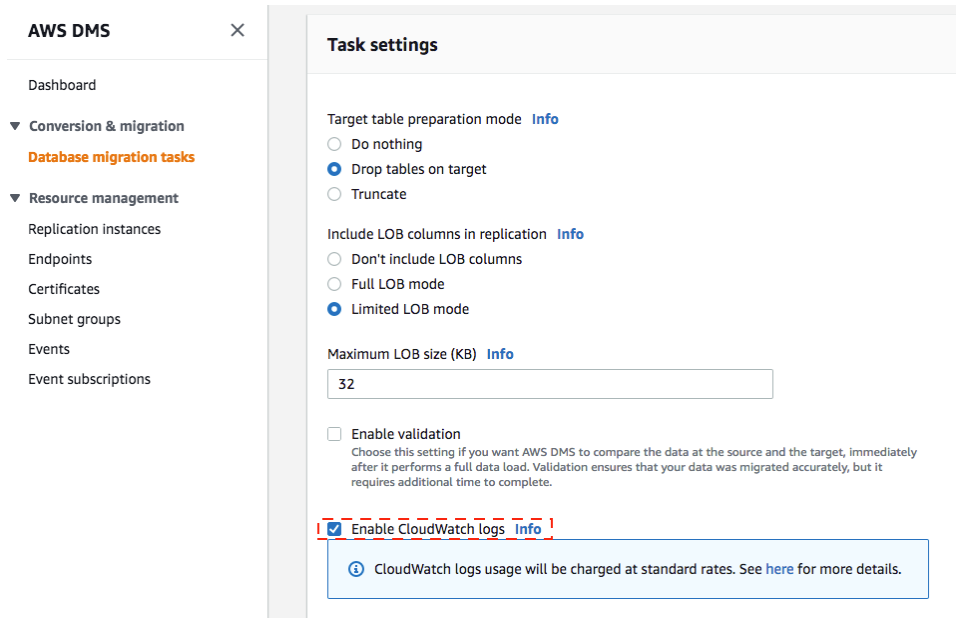


2. Click **Create Task**.
   a. Type an easily recognized **Task name** e.g. **dms-full-dump-task**

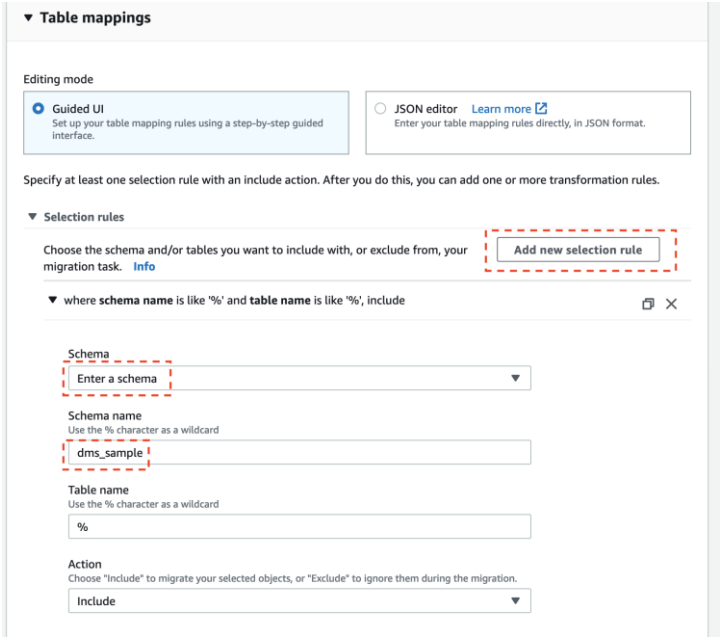
b. Select your **Replication instance** from drop down.
c. Select your **Source endpoint** from drop down.
d. Select your **Target endpoint** from drop down.
e. For, **Migration type** choose **Migrate existing data**.
f. Select **Start task on create**



g. Under **Task Settings,** select the **Enable CloudWatch logs** check box.

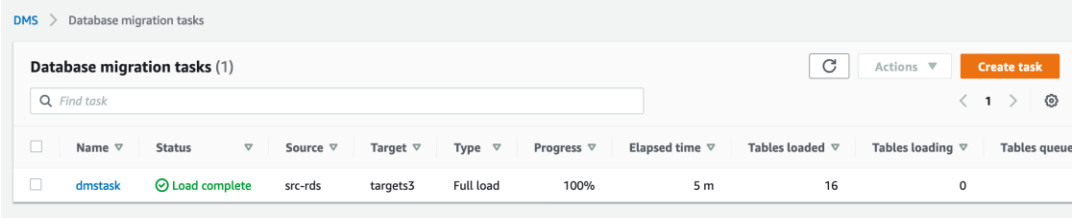


h. Go to **Table Mappings**.
i. Click on **Add new selection rule** and select **Enter a Schema** in **Schema** field.
j. For Schema name, type **dms_sample** and keep the settings for the remaining fields

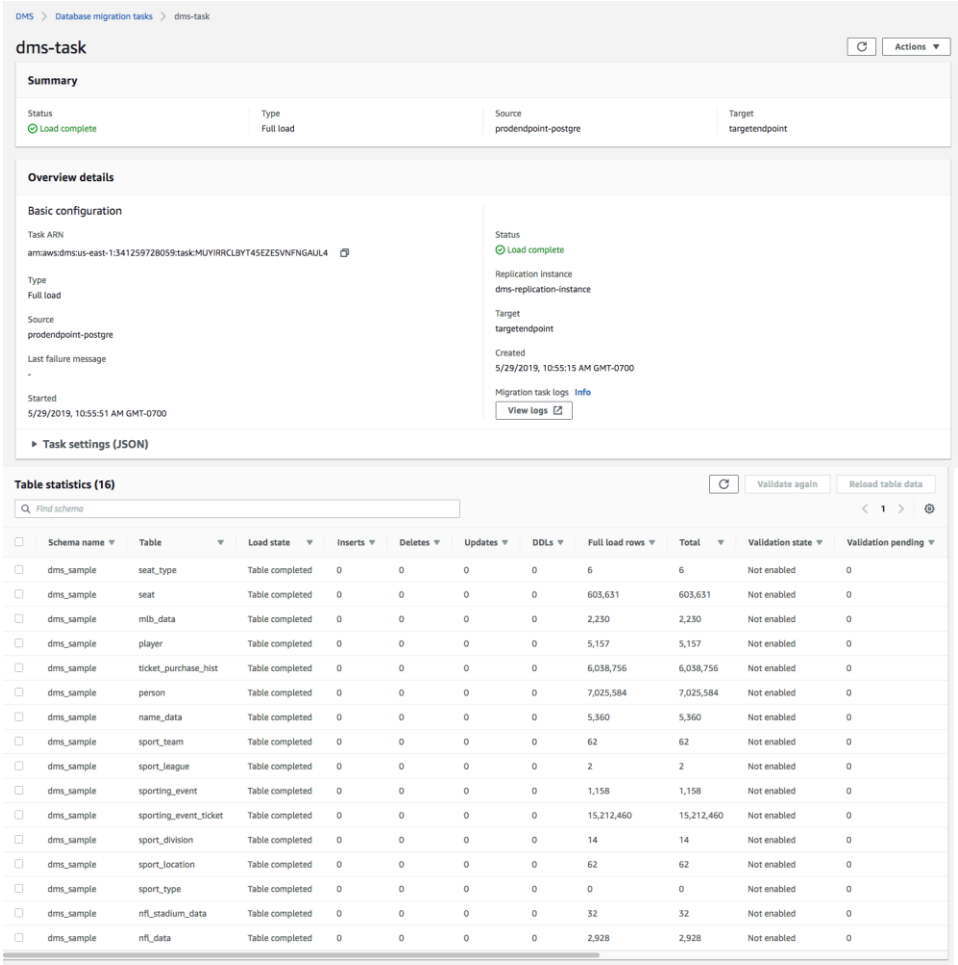


3. Click **Create task**. Your task is created and starts automatically. (Note: The complete creation and data extraction process takes around 5 minutes.)

4. Once complete, the console displays 100% complete.

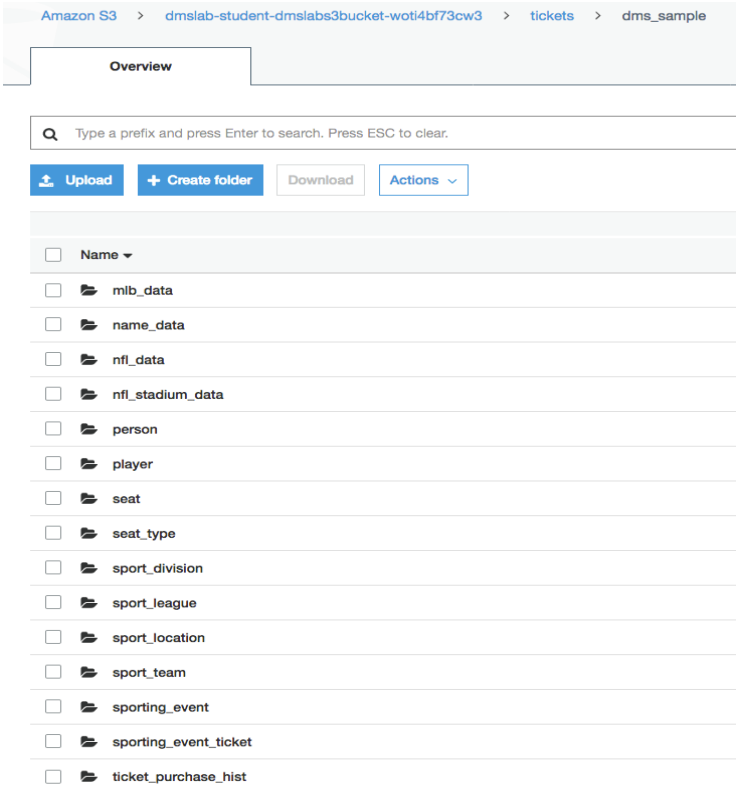


5. Select your task and explore the summary. Under **Table** statistics tab you can review all table information loaded in S3 from RDS by DMS



6. Open the S3 console to view the data that was copied by DMS: https://s3.console.aws.amazon.com/s3/home?region=us-east-1#

7. Click on the bucket used as the DMS target and navigate to **/tickets/dms_sample/** to view the loaded tables, one folder per table



8. Download one of the files:
    a. Navigate further to **mlb_data/LOAD00000001.csv**, select the check box next to the file name and click **Download** in the pop-up window.
    b. Click **Save File**.
    c. Open the file.

    You will notice that the file contains the column headers in the first row as requested by the “addColumnName=true” connection attribute we included when we created the s3 target endpoint. Note that column names are included in the file in the first row.

	A	B	C	D	E
1	id	sport_team_id	last_name	first_name	full_name
2	1	131	Adam Loewen	Adam	Loewen
3	11	131	A.J. Pollock	A.J.	Pollock
4	21	131	Alex Sanabia	Alex	Sanabia
5	31	131	Andrew Chafin	Andrew	Chafin
6	41	131	Andy Marte	Andy	Marte
7	51	131	Archie Bradley	Archie	Bradley
8	61	131	Ben Francisco	Ben	Francisco
9	71	131	Braden Shipley	Braden	Shipley
10	81	131	Bradin Hagens	Bradin	Hagens
11	91	131	Brandon Drury	Brandon	Drury
12	101	131	Brett Jackson	Brett	Jackson

You may notice that the primary key column was loaded in scientific notation:

LOAD00000001.csv - Notepad

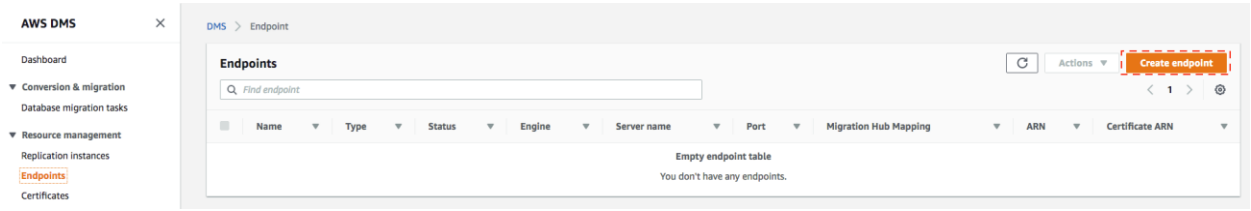
File Edit Format View Help

```
mlb_id,mlb_name,mlb_pos,mlb_team,mlb_team_long,bats,throws,birth_year,bp
+5.0656000000000000e+05,Alexi Amarista,3B,SD,San Diego Padres,L,R,1989,+
+4.5821000000000000e+05,Alexi Casilla,2B,TB,Tampa Bay Rays,S,R,1984,+4.5
+4.6839600000000000e+05,Alexi Ogando,P,ATL,Atlanta Braves,R,R,1983,+4.99
+4.6968600000000000e+05,Alfredo Aceves,P,NYY,New York Yankees,R,R,1982,+
+4.5162800000000000e+05,Alfredo Figaro,P,TEX,Texas Rangers,R,R,1984,+5.2
+5.5405400000000000e+05,Alfredo Gonzalez,C,CWS,Chicago White Sox,R,R,199
+5.0124500000000000e+05,Alfredo Marte,LF,BAL,Baltimore Orioles,R,R,1989,
+4.3058000000000000e+05,Alfredo Simon,P,CIN,Cincinnati Reds,R,R,1981,+4.
+4.5537800000000000e+05,Ali Solis,C,LAD,Los Angeles Dodgers,R,R,1987,+5.
+4.8885200000000000e+05,Allan Dykstra,1B,TB,Tampa Bay Rays,L,R,1987,+5.7
+5.0180000000000000e+05,Allen Craig,1B,BOS,Boston Red Sox,R,R,1984,+5.16
```

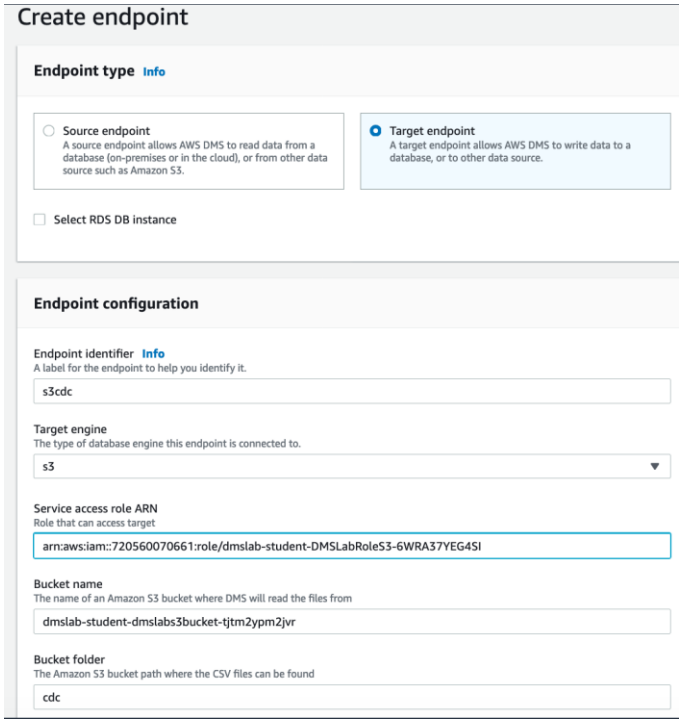
This is due to the tables at the source having primary key as **double precision**. Keep in mind that DMS allows you to perform additional transformations, for example type casting at load time. Here we will proceed without making any further modifications.

(Optional) Create a DMS endpoint to perform ongoing replication

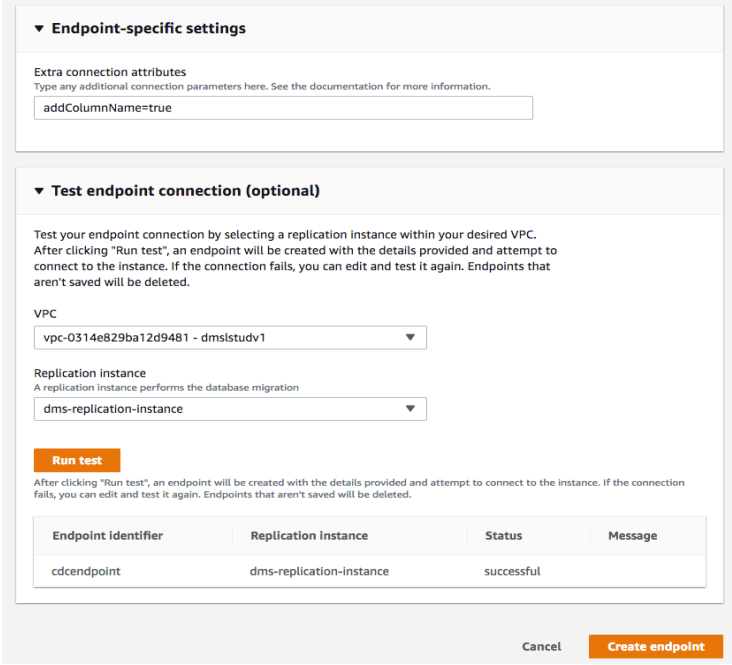
1. Navigate to the DMS console: https://console.aws.amazon.com/dms/v2/home?region=us-east-1#dashboard and select **Endpoints**:



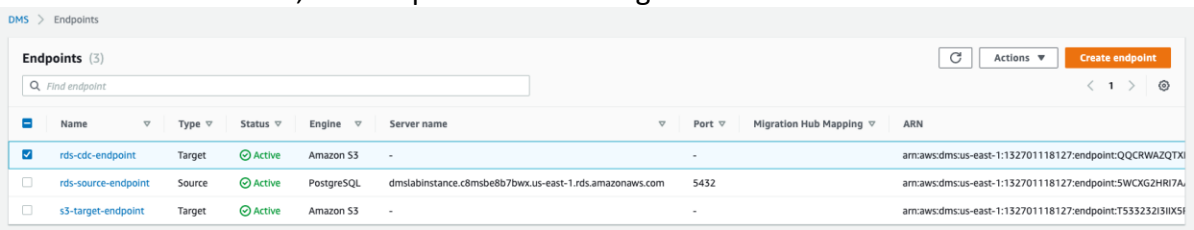
2. Click **Create endpoint**.
    a. For **Endpoint type**, select **Target**
    b. For **Endpoint identifier**, type **rds-cdc-endpoint**
    c. For **Target engine**, choose **s3**.
    d. For Service access role ARN, paste the **DMSLabRoleS3** number noted earlier
    e. For Bucket name, paste the **S3 Bucket Name** noted earlier
    f. For **Bucket folder**, type **cdc**.



g. Click **Endpoint-specific settings** to expand the section.
h. In the **Extra connection attributes** box, type **addColumnName=true** to include column names in the files in the S3 bucket.
i. Expand the **Test endpoint connection (optional)** section, and choose your **dmslstudv1** name on the VPC drop-down list.
j. Click Run test. This step tests connectivity to the source database system. If successful, the message “Connection tested successfully” appears.

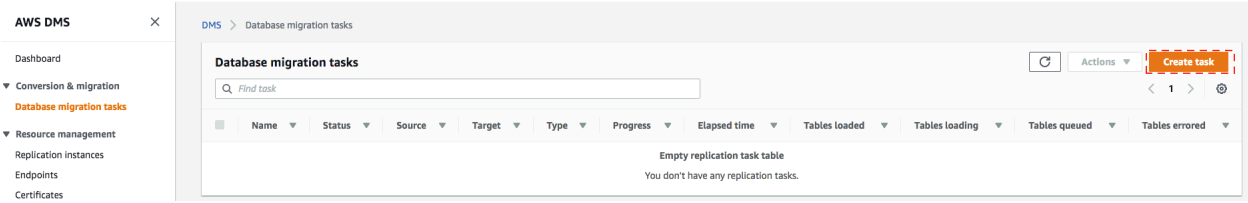


3. Click **Create endpoint**.
4. When available, the endpoint status changes to active.

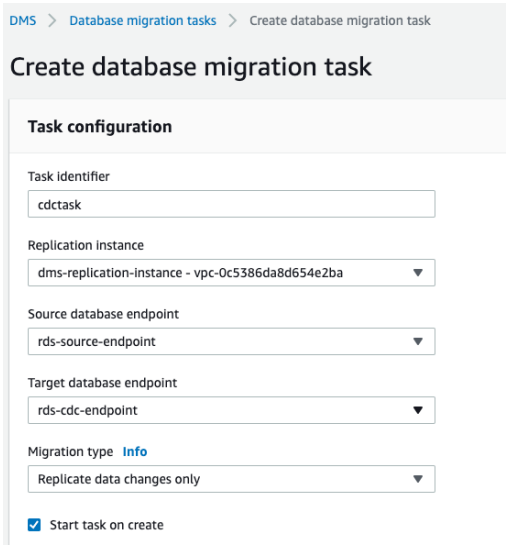


(Optional) Create a task to perform ongoing replication

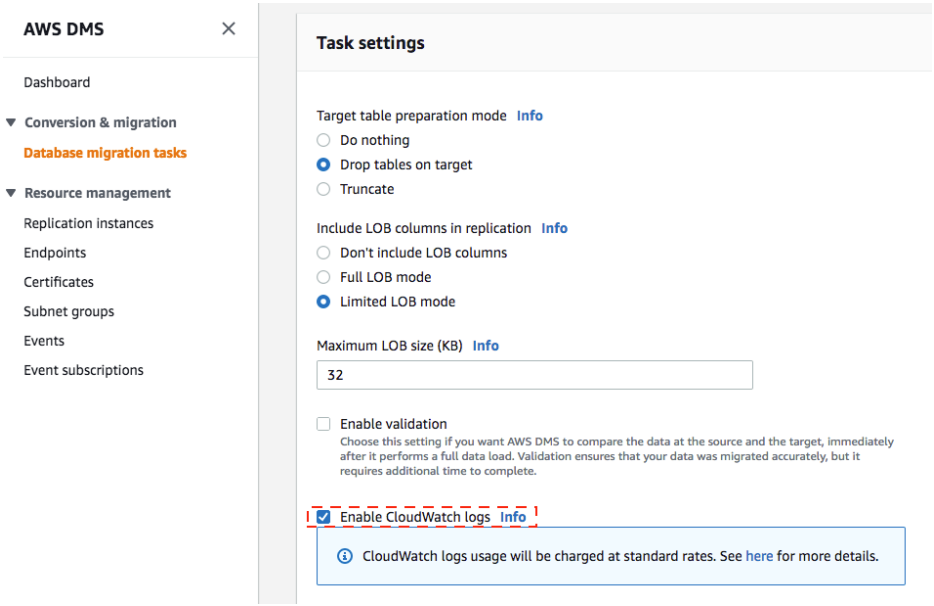
1. Navigate to the DMS console: https://console.aws.amazon.com/dms/v2/home?region=us-east-1#dashboard and select **Database Migration Tasks**.



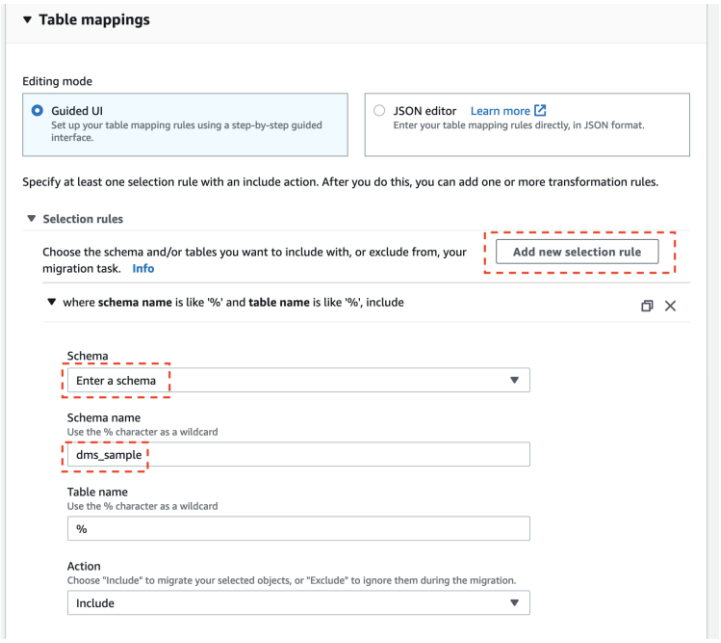
2. Click **Create Task**.
    a. Type **cdctask** as **Task Identifier**
    b. Select your **Replication instance**.
    c. Select your **Source endpoint**.
    d. Select **Target endpoint** as **rds-cdc-endpoint** created in the previous section.
    e. For **Migration type**, choose **Replicate data changes only**.
    f. Select the Start task on create check box.



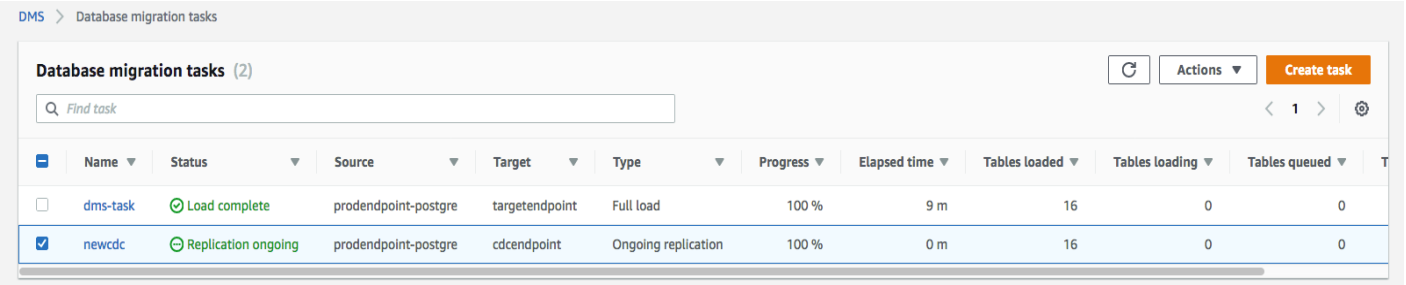
g. In **Task Settings**, Select the **Enable CloudWatch logs** check box. Do not enable the validation.



h. Go to **Table Mappings**.
i. Click on **Add new selection rule** and select **Enter a Schema** in Schema field.
j. For **Schema name**, type **dms_sample** and keep the values in the remaining fields

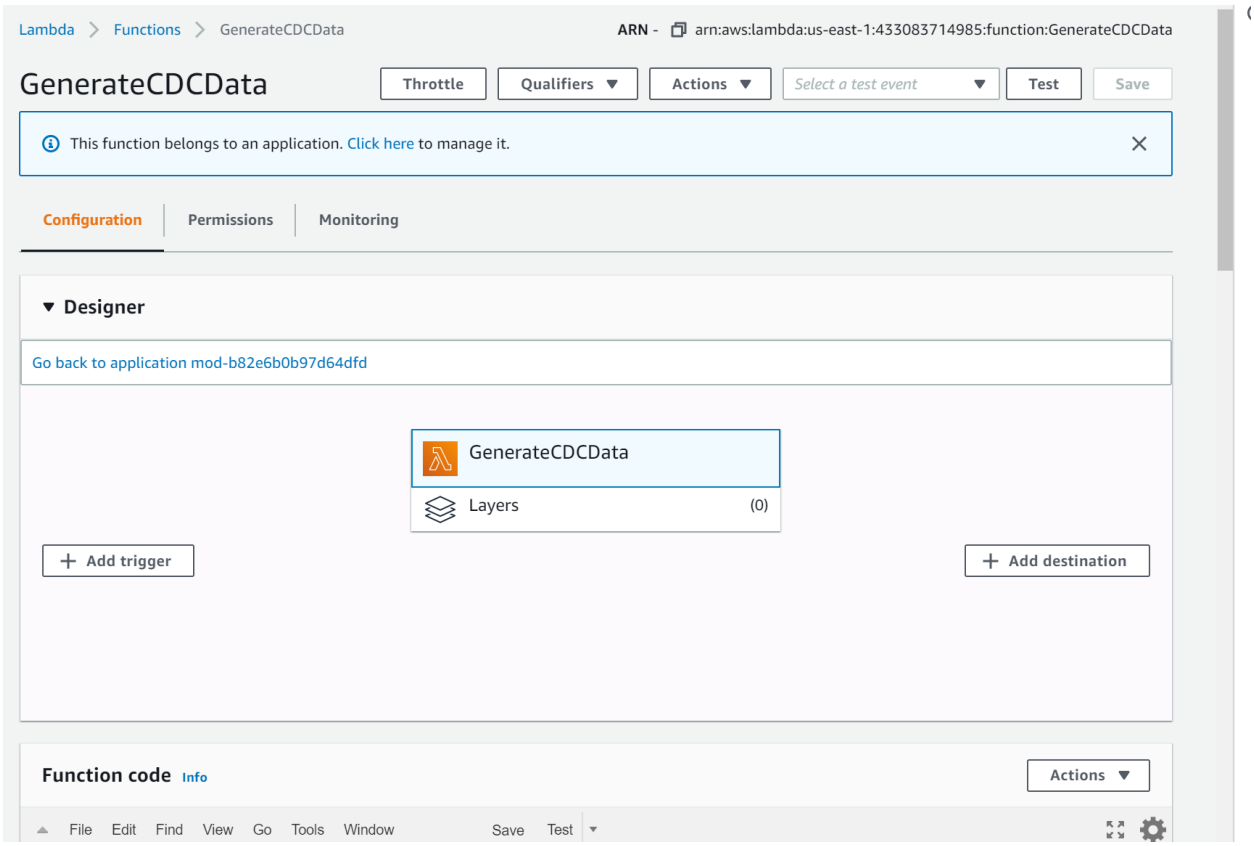


3. Click Create task. Your task is created and starts automatically. You can see status as **ongoing replication**, after couple of minutes.

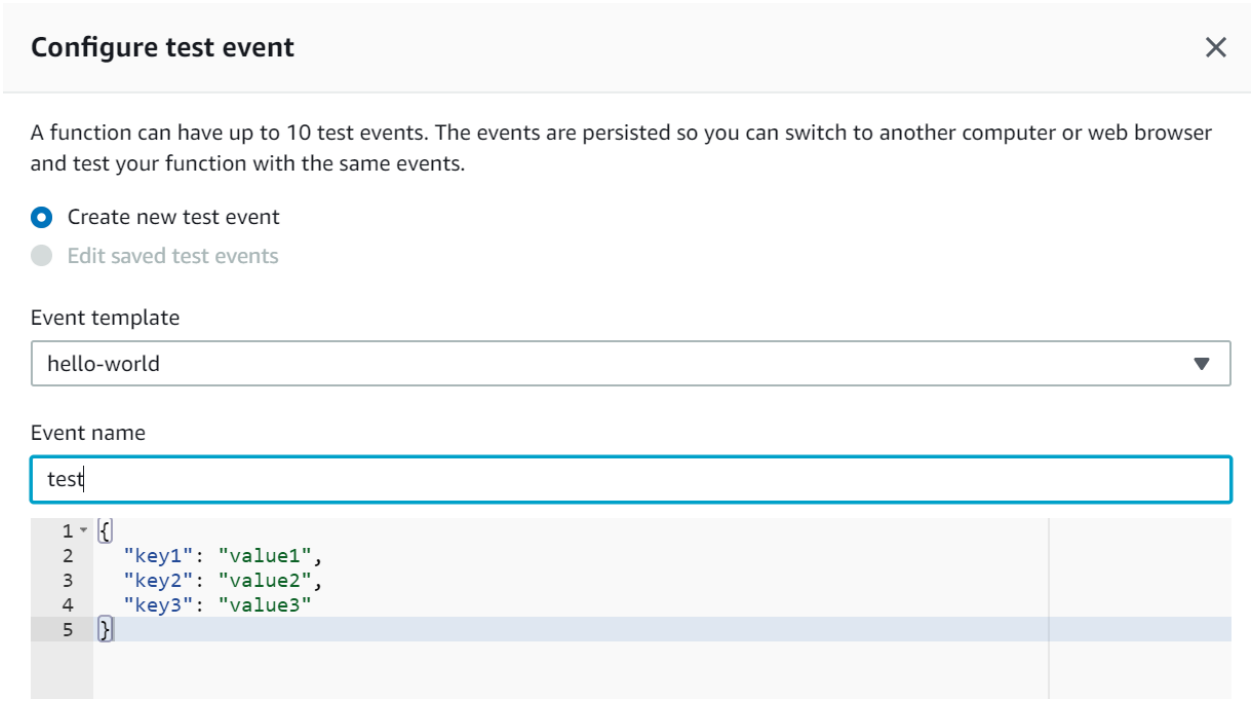


Once complete, the console displays 100% complete.

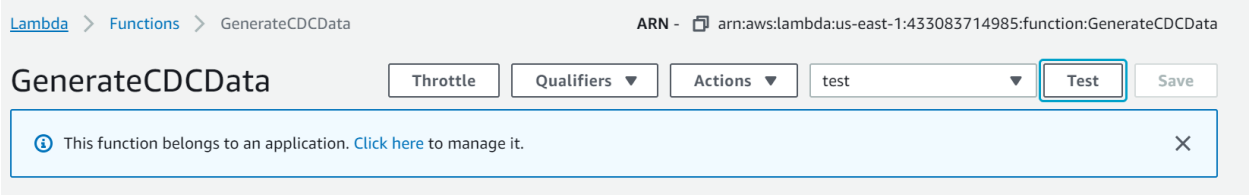
4. To generate CDC activity which above migration task will capture, we provisioned a Lambda function **GenerateCDCData** that will generate some new data in the source database. Start a new browser tab and navigate to https://console.aws.amazon.com/lambda/home?region=us-east-1#/functions/GenerateCDCData?tab=configuration.



We will be invoking the Lambda function via a test event. Click **Test** and in the popup window, enter **test** under **event name** and click **Create** to create a test event:



Now with a test event created, click **Test** again (feel free to click it a few times) to generate some changes in the source database:



You may need to wait 5 to 10 minutes for CDC data to first reflect in your RDS postgres database and then picked up by DMS CDC migration task.

5. Once the CDC Data gets replicated, you can navigate to CDC task details, and under **Table statistics** tab review the details, as shown below:

Note: In case you see DMS CDC task in fail/error status. Make sure your replication instance version is 3.3.1 and it is large enough (dms.t2.medium or above) to run CDC replication task

Table statistics (16)

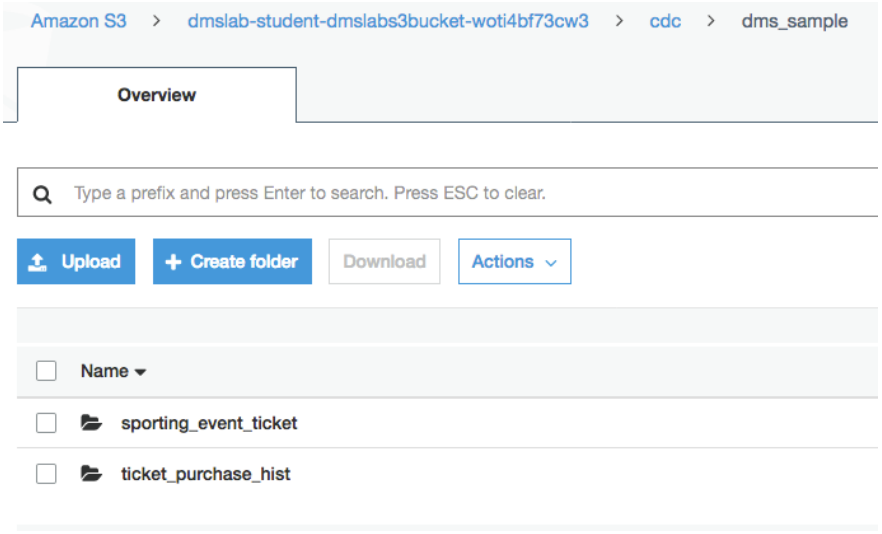
Validate again | Reload table data

Find schema

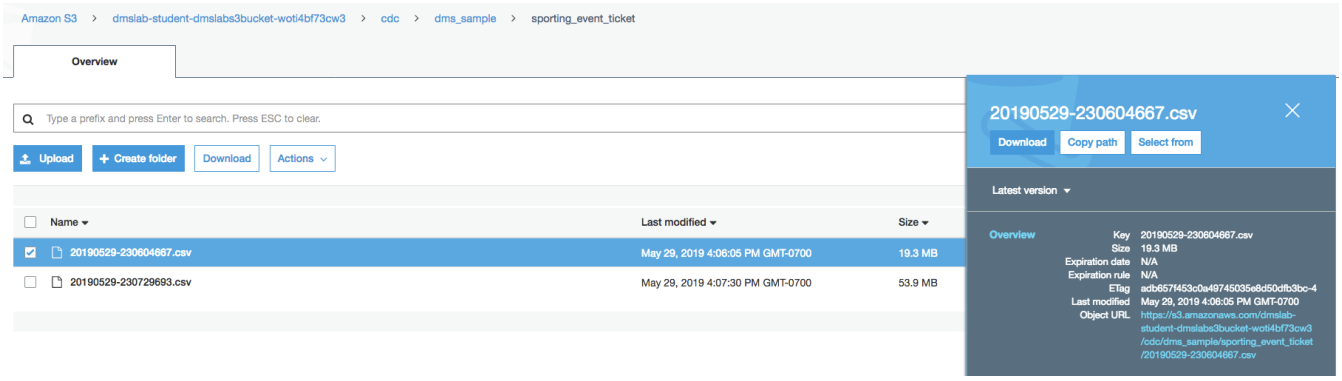
Schema name	Table	Load state	Inserts	Deletes	Updates	DDLs	Full load rows	Total	Validation state	Validation pending
dms_sample	seat_type	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	seat	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	mlb_data	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	player	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	ticket_purchase_hist	Table completed	11,002	0	0	0	0	11,002	Not enabled	0
dms_sample	person	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	name_data	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sport_team	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sport_league	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sporting_event	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sporting_event_ticket	Table completed	0	0	11,002	0	0	11,002	Not enabled	0
dms_sample	sport_division	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sport_location	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	sport_type	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	nfl_stadium_data	Table completed	0	0	0	0	0	0	Not enabled	0
dms_sample	nfl_data	Table completed	0	0	0	0	0	0	Not enabled	0

6. Open the S3 console to view the data that was copied by DMS: https://s3.console.aws.amazon.com/s3/home?region=us-east-1#

7. Click on the bucket used as the DMS target and navigate to **/cdc/dms_sample/** to view the loaded tables, one folder per table



8. Download one of the files:
    a. Select the check box next to the object name and click Download in the pop-up window.
    b. Click **Save File**.
    c. Open the file.
    You will notice that the file contains the column headers in the first row as requested by the **addColumnName=true** connection attribute we included when we created the s3 target endpoint.



Note that file name has a timestamp. You can see the header is included and the operation column is added at the beginning of each row. The file below shows updates (U) to the table along with the values after the update. Inserts (I) show data after the insert and Deletes (D) show data before the delete.

	A	B	C	D	E	F	G	H	I	J
	Op	id	sporting_eve	sport_location_id	seat_level	seat_section	seat_row	seat	ticketholder_id	ticket_price
	U	145192591	3931	4	2	10	A	2	2898028	98
	U	145192601	3931	4	2	10	A	1	2898028	98
	U	145192581	3931	4	2	10	A	3	2898028	98
	U	145192501	3931	4	2	10	B	1	2898028	98
	U	145187751	3931	4	2	13	B	2	2898028	49
	U	145187741	3931	4	2	13	B	3	2898028	49
	U	145187721	3931	4	2	13	C	2	2898028	49
	U	145187711	3931	4	2	13	C	3	2898028	49
	U	145187731	3931	4	2	13	C	1	2898028	49
	U	145187701	3931	4	2	14	A	1	2898028	49
	U	145187681	3931	4	2	14	A	3	2898028	49
	U	145187691	3931	4	2	14	A	2	2898028	49
	U	145187471	3931	4	2	14	B	3	2898028	49
	U	145187671	3931	4	2	14	B	1	2898028	49
	U	145187481	3931	4	2	14	B	2	2898028	49
	U	145187451	3931	4	2	14	C	2	2898028	49
	U	145187461	3931	4	2	14	C	1	2898028	49
	U	145190341	3931	4	2	14	C	3	2898028	49
	U	145183201	3931	4	2	15	A	4	2898028	49
	U	145179691	3931	4	2	15	A	1	2898028	49
	U	145179661	3931	4	2	15	A	4	2898028	49
	U	145179671	3931	4	2	15	A	3	2898028	49
	U	145179681	3931	4	2	15	A	2	2898028	49
	U	145190321	3931	4	2	15	A	2	2898028	49